第 41 卷 第 3 期 天 文 学 进 展 Vol. 41, No. 3
2023 年 9 月 PROGRESS IN ASTRONOMY Sept., 2023

doi: 10.3969/j.issn.1000-8349.2023.03.09

基于机器学习的单脉冲搜索候选体识别对 FAST 观测 CRAFTS 数据的应用研究

张 彬[1,2,3,4]，游善平[1,3,4]，谢晓尧[1,3,4]，于徐红[1,3,4]，梁 楠[1,3,4]

(1. 贵州师范大学 贵州省信息与计算科学重点实验室/网络空间安全学院，贵阳 550001； 2. 贵州师范大学 数学科学学院，贵阳 550001； 3. 中国科学院-贵州师范大学 FAST 早期科学数据中心，贵阳 550001； 4. 中国天眼联合研究中心贵州师范大学分中心，贵阳 550001)

摘要： 单脉冲搜索作为脉冲星探测的有力工具，在探测旋转射电暂现源以及快速射电暴中扮演着重要角色。为了从海量的射电巡天数据中快速筛选出最有价值的单脉冲搜索候选体，候选体识别已经从早期启发式阈值判断发展到基于机器学习自动识别。对于 FAST 观测，研究了基于机器学习的单脉冲搜索候选体识别应用到 CRAFTS (the commensal radio astronomy FAST survey) 超宽带脉冲星数据的性能表现。在评估过程中，使用单脉冲事件组识别 (SPEGID) 和单脉冲搜索器 (SPS) 两类自动识别方法，通过 7 种不同机器学习分类器对 CRAFTS 基准数据集产生的单脉冲搜索候选体进行自动识别；作为对比，也使用了启发式阈值判断的方法 (RRATtrap 和 Clusterrank)。结果表明，SPEGID 具有最好的性能表现 (最高的 F1-score 值 95.1%、次高的召回率 95.4%、最低的假阳性率 4.7%)，SPS 具有最快的筛选速度 (平均每小时筛选 4 010 个候选体)。通过对比分析结果，探讨了如何基于 FAST 观测数据开展高效的单脉冲搜索候选体识别。

关 键 词： 单脉冲搜索；候选体识别；机器学习；脉冲星；FAST；CRAFTS

中图分类号： P111.44 **文献标识码：** A

1 引 言

脉冲星搜索方法主要分为周期性搜索和单脉冲搜索两大类[1]。周期性搜索通过快速傅里叶变换 (FFT) 将时间序列转化到频域以识别周期性信号[2]。传统上主要通过周期性搜索来探测脉冲星，这是利用脉冲星信号固有的周期性来实现。单脉冲搜索主要寻找强的、非周期

收稿日期： 2022-09-06； **修回日期：** 2023-03-06

资助项目： 中国科学院天文大科学研究中心 FAST 重大成果培育项目 (FAST[2019sr04])；贵州省科学技术基金 (黔科合基础-ZK[2021] 重点 020，黔科合 J 字 LKS[2010]38 号)

通讯作者： 梁楠，liangn@bnu.edu.cn

的脉冲，非常适合发现周期性搜索中无法发现的孤立爆发[3]；应用单脉冲搜索方法导致了旋转射电暂现源 (rotating radio transients, RRATs) 和快速射电暴 (fast radio bursts, FRB) 的发现。2006 年，McLaughlin 等人[4]首先发现了 RRATs，被认为是一种特殊类型的间歇脉冲星。2007 年，Lorimer 等人[5]在帕克斯多波束脉冲星巡天 (Parkes multibeam pulsar survey, PMPS) 观测数据中发现了第一例 FRB。

自 2003 年 Cordes 和 McLaughlin[3] 首次提出单脉冲搜索探测脉冲星以来，单脉冲搜索的应用产生了海量候选体，为了从射电巡天数据中快速筛选出最有价值的候选体，基于特定的脉冲星巡天数据面临的候选体识别问题，相继提出了不同的单脉冲搜索候选体识别方法[6–11]。单脉冲搜索候选体识别已经从早期启发式阈值判断发展到基于机器学习 (machine learning, ML) 自动识别[6]。启发式阈值判断的识别方法主要利用脉冲星所具有的启发式特性来引导搜寻，筛选出最有价值的单脉冲搜索候选体。例如：Deneva 等人[7]通过对阿雷西博 L 波段馈源阵列脉冲星巡天 (pulsar Arecibo L-band feed array survey, PALFA) 单脉冲搜索，发现了 7 颗新脉冲星。Keane 等人[8]在 PMPS 中发现了 10 颗 RRATs。Burke-Spolaor 等人[9]在高时间分辨率宇宙脉冲星巡天 (high time resolution universer survey, HTRU) 观测数据中发现了 11 颗 RRATs。2015 年，Karako-Argaman 等人[10]设计用于探测脉冲星和 RRATs 的工具 RRATtrap，根据候选体与设定的规则符合程度分配数值分数，通过只检查超过给定阈值的候选体，区分脉冲星与射频干扰产生的候选体，在绿岸望远镜 350 MHz 漂移扫描巡天 (Green Bank telescope 350-MHz drift-scan survey, GBT350Drift) 和绿岸北天区巡天 (Green Bank north celestial cap survey, GBNCC) 的观测数据中发现 21 颗 RRATs。2016 年，Deneva 等人[11]开发 Clusterrank 工具，通过量化候选体色散与信噪比曲线与 Cordes 和 McLaughlin 预测的理论曲线[3] 符合程度，评判候选体是脉冲星的可能性，在阿雷西博 327 MHz 漂移脉冲星巡天数据 (Arecibo 327 MHz drift pulsar survey, AO327) 中发现 14 颗脉冲星和 8 颗 RRATs。启发式阈值判断方法主要根据脉冲星的特性构建启发式规则，并没有针对射频干扰构造规则进行过滤，往往会产生大量的虚假候选体，难以适应大规模、海量的数据处理。

目前，人工智能相关技术已广泛应用在周期性搜索候选体识别任务中[12–15]，而在单脉冲搜索候选体识别领域的研究相对较少。随着脉冲星巡天设备产生的候选体数量呈指数增长，仅依赖人工识别筛选已不能满足数据的时效需求，机器学习等人工智能技术已经开始逐渐运用到单脉冲搜索候选体识别研究领域。基于机器学习的单脉冲搜索候选体识别方法 (以下简称“机器学习识别方法”)，利用脉冲星与射频干扰固有的特性开发特征工程，构建强有力的特征以最大限度区分脉冲星与射频干扰，通过机器学习分类器对候选体进行自动识别。2016 年，Devine 等人[16]首次将机器学习应用于单脉冲搜索候选体识别，实现了自动化筛选候选体；2018 年，同组的 Pang 等人[17] 提出了单脉冲事件组识别 (single-pulse event group identification, SPEGID)，构造 18 个特征描述聚合产生的单脉冲事件组 (single-pulse event group, SPEGs)，结合机器学习分类器对 PALFA 观测数据进行自动识别；随后，SPEGID 特征工程被拓展到 23 个[18]，并应用到 GBTDrift。另外一方面，2018 年，Michilli 等人[19] 设计了单脉冲搜索器 (single-pulse searcher, SPS)，用 5 个特征描

述聚合产生的 SPEGs，并通过机器学习分类器区分低频阵列 (low frequency array, LOFAR) 全天空巡天数据 (tied-array all-sky survey, LOTAAS) 强干扰环境下的脉冲星与射频干扰。

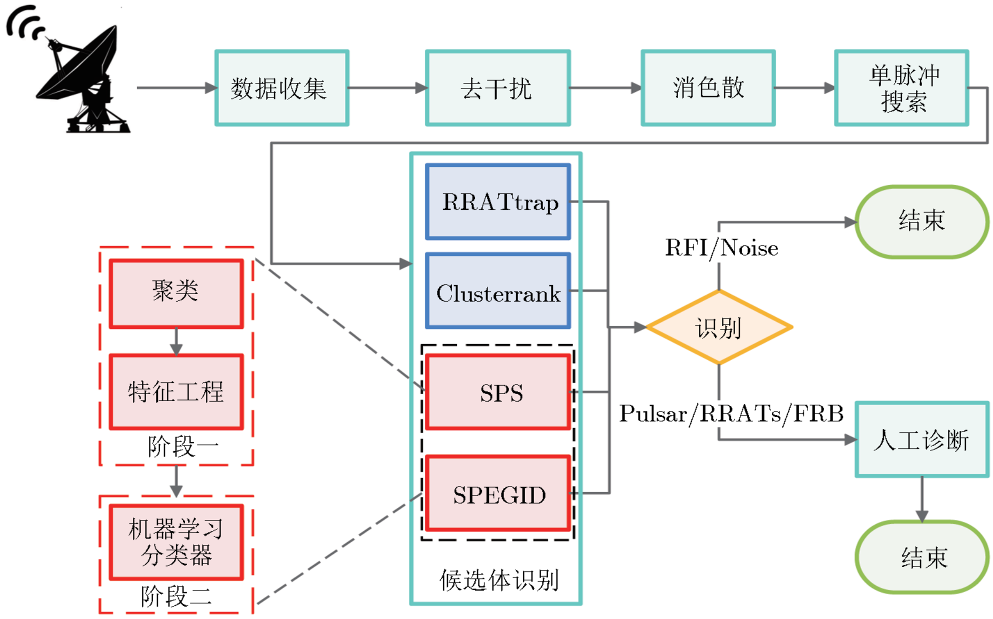
500 m 口径球面射电望远镜 (Five-hundred-meter Aperture Spherical radio Telescope, FAST) 是目前世界上最灵敏的单口径射电望远镜。FAST“多科学目标同时扫描巡天” (the commensal radio astronomy FAST survey, CRAFTS) 同时使用多个数字终端采集脉冲星、中子氢、分子谱线、暂现源、FRB 等多科学目标观测数据[20]。据估计，CRAFTS 脉冲星搜索在每次 24 h 巡天扫描中会产生数万到数十万个脉冲星候选体[20]，通过人工诊断发现，这些候选体绝大多数是射频干扰或宇宙噪音引起的虚假候选体[21]。对 FAST 这种大规模射电天文观测数据进行单脉冲搜索寻找新脉冲星，必须快速找到具有科学价值的单脉冲搜索候选体并对其进行优先存储，以避免累积延迟，并使用稳健的单脉冲搜索候选体识别方法准确高效地区分出脉冲星与射频干扰。

2017 年 8 月至 2018 年 5 月，FAST 使用超宽带接收机 (270 ~ 1 620 MHz) 运行漂移扫描巡天模式，快速连续地观测天空中的多个区域。期间共收集 2 760 h 的脉冲星巡天数据，共计 317 497 个数据文件，称为 CRAFTS 超宽带脉冲星巡天数据 (以下简称“CRAFTS 数据”)，存储在中国科学院国家天文台-贵州师范大学 FAST 早期科学数据中心。本研究工作主要利用机器学习识别方法对 CRAFTS 数据产生的单脉冲搜索候选体的性能表现进行评估，以寻求快速高效地筛选出候选体的解决方案。本文结构安排如下：第 2 章介绍基于机器学习的单脉冲搜索候选体识别方法基本理论；第 3 章使用机器学习识别方法 (SPEGID 和 SPS) 和启发式阈值判断识别方法 (RRATtrap 和 Clusterrank) 对 CRAFTS 单脉冲搜索产生的候选体基准数据集进行测试，并对不同识别方法性能表现以及速度进行对比分析；第 4 章对全文进行总结和讨论。

2 基于机器学习的单脉冲搜索候选体识别方法基本理论

利用射电天文观测数据探测脉冲星通常分为五个阶段[1]：数据收集、去干扰、消色散、周期性搜索或单脉冲搜索、人工诊断。第一阶段，射电望远镜终端收集到的射电信号以电压时间序列的形式存储；第二阶段，去干扰是消除或减轻射频干扰对搜索结果的影响；第三阶段，消色散是消除与频率有关的延迟效应的影响；第四阶段，使用周期性搜索或单脉冲搜索筛选出观测数据中的脉冲星候选体；第五阶段，对每个脉冲星候选体进行人工诊断，以确定其真实性。目前，FAST 观测数据主要使用并行化实现的 PRESTO (pulsar exploration and search toolkit) 开展脉冲星搜索研究[22]。基于 PRESTO 的单脉冲搜索方法探测天体物理信号的数据流程图主要包括如下步骤：数据收集、去干扰、消色散、单脉冲搜索、候选体识别和人工诊断 (如图 1 所示)。

在图 1 候选体识别模块中，同时列出了启发式阈值判断识别方法 (RRATtrap 和 Clusterrank) 以及机器学习识别方法 (SPEGID 和 SPS)。表 1 给出了以上 4 类识别方法的基本信息。为了比较不同候选体识别方法区分脉冲星与非脉冲星候选体的能力，可以通过



注：虚线框放大区域展示机器学习识别方法两阶段数据处理流程图。

图 1　基于 PRESTO 的单脉冲搜索方法探测天体物理信号流程图

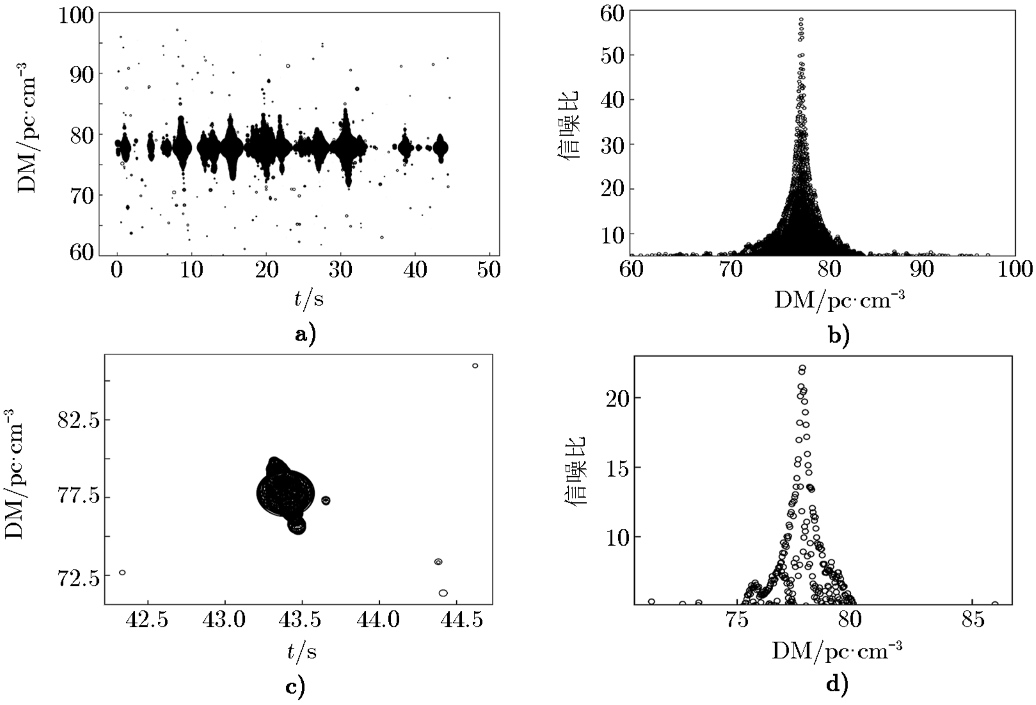
表 1　4 类单脉冲搜索候选体识别方法的相关信息

类型	识别方法	相关文献	观测数据集	性能表现	
				召回率	假阳性率
启发式阈值判断	RRATtrap	[10]	GBT350Drift GBNCC	0.8	0.09
	Clusterrank	[11]	AO327	0.7	0.07
机器学习	SPEGID	[17]	PALFA	0.956	0.02
		[18]	GBTDrift	0.942	0.02
	SPS	[19]	LOTAAS	0.986	0.01

评价指标①对性能表现进行量化。候选体识别任务中，我们希望识别方法尽可能准确地识别所有脉冲星候选体的同时，最大程度减少虚假候选体的产生。因此，衡量候选体识别方法的性能表现最主要的评价指标是召回率和假阳性率[10]：召回率量化识别方法正确识别数据集中脉冲星候选体的比例，假阳性率量化识别方法产生虚假候选体的比例。最佳识别方法具有高召回率以及低假阳性率。F1-score 及 G-mean 综合评估识别方法正确识别脉冲星以及产生虚假候选体的能力，最均衡的识别方法具有高 F1-score 和 G-mean 值。从表 1 可知，机器学习识别方法相比启发式阈值判断具有高召回率和低假阳性率。

①常用的评价指标有：准确率，正确分类脉冲星和非脉冲星的数量占训练实例的百分比；查准率，正确分类为脉冲星与被分类为脉冲星的实例数量比值；召回率，正确分类为脉冲星与真实为脉冲星的实例数量比值；假阳性率，被误判为脉冲星占非脉冲星的实例总数百分比；F1-score，查准率和召回率的加权调和平均值；G-mean，召回率和特异度 (正确分类为非脉冲星与真实为非脉冲星的实例数量比值) 乘积的算术平方根。

单脉冲搜索通常会为每个观测生成一个或多个诊断图，从诊断图中分离出脉冲星的过程被称为候选体识别。图 2 给出通过 PRESTO 对 CRAFTS 数据单脉冲搜索探测到的 PSR B0540+23 结果诊断图。基于 PRESTO 的单脉冲搜索候选体识别方法，通过处理单脉冲搜索 Single_pulse_search.py 结果文件 (记录每个单脉冲事件的色散值、脉冲到达时间、信噪比、脉冲宽度)，实现候选体的识别与分类任务。通常假设脉冲星信号具有显著特征，可以在搜索过程中从大量虚假探测中脱颖而出，单脉冲搜索候选体识别方法被设计寻找这些隐藏的特征[6]。由于脉冲星信号通常与出现在一定色散 (dispersion measure, DM) 范围大约相同时间的单脉冲事件组紧密相关。单脉冲搜索候选体识别方法一般先通过聚类算法将出现在一定 DM 和时间阈值范围内的单脉冲事件 (single-pulse event, SPE) 聚合成单脉冲事件组 (SPEGs)；再构造区分脉冲星与非脉冲星候选体的启发式规则或开展特征工程；最后应用构建的规则或机器学习分类器对 SPEGs 进行识别，进而实现候选体的分类任务。



注：a) 时间与试验 DM 的散点图，每个散点对应一个单脉冲事件 (SPE)，大小与 SPE 的信噪比成正比；b) 试验 DM 对应的信噪比大小；c) 一组 SPE 聚合为一个单脉冲事件组 (SPEGs) 时间与 DM 空间分布；d) 一个 SPEGs 对应的 DM 与信噪比空间分布。

图 2　通过 PRESTO 对 CRAFTS 数据单脉冲搜索探测到 PSR B0540+23 结果诊断图

机器学习识别方法一般分两个阶段对单脉冲搜索候选体自动识别与分类 (如图 1 虚线框放大区域)。第一阶段通过聚类算法将相关单脉冲事件 (SPE) 聚合成单脉冲事件组 (SPEGs) 后开发特征工程。第二阶段结合机器学习算法，创建一个完全标记的特征数据集训练多

种机器学习分类器。在第一阶段，SPEGID 方法结合具有噪声基于密度的空间聚类算法 (DBSCAN)[23] 将相关 SPE 聚合成任意形状 SPEGs；随后构造 18 个特征①描述 SPEGs。SPS 方法采用 Friends-of-Friends 聚类算法[24] 将相邻时间和 DM 阈值内的 SPE 聚合成 SPEGs；通过 5 个特征②对 SPEGs 进行统计意义上的建模来描述 SPEGs。在第二阶段，SPEGID 方法选择数据挖掘软件 WEKA 实现的 6 种分类器；最后使用最佳分类器 (RandomForest) 对未标记的观测数据进行自动识别和分类。SPS 方法选择最佳分类器高斯海灵格快速决策树 (GH-VFDT)[25] 对 SPEGs 进行识别；并根据空间信息对标记为脉冲星 SPEGs 进一步过滤，最终生成候选诊断图以供人工诊断。

为了实现机器学习分类器对单脉冲搜索候选体的自动识别与分类，要求分类器能够从训练集中获得“脉冲星”的一般模式，这是监督学习的一个应用，是指从标记为脉冲星与非脉冲星的训练集中推断出区分脉冲星与射频干扰的目标函数[26]，该函数可以根据观测数据特征值将其准确地映射到对应的类别 (脉冲星、非脉冲星)。另外，在海量射电天文观测数据中，绝大多数是射频干扰或宇宙噪音引起的无用数据，仅有极少数探测到脉冲星信号，因此射电天文观测数据存在严重的类别不平衡[6]。而机器学习分类器在类别不平衡的数据集训练时，分类器通常会对多数类别 (非脉冲星) 进行“过度训练”，导致训练的分类器对新的观测数据进行分类时，分类结果会偏向多数类别，致使感兴趣的少数类别 (脉冲星) 出现大量误判[27]。为了缓解机器学习分类器在不平衡数据集性能表现较差的问题，须对基准数据集进行不平衡处理。之前基于机器学习的单脉冲搜索候选体识别方法的研究表明[16, 17]，SMOTE (合成少数群体过采样技术)[28] 在数据不平衡的处理上优于其他方法。

3 机器学习识别方法对 CRAFTS 数据的应用和对比分析

在本研究工作中，我们通过测试四类单脉冲搜索候选体识别方法对 CRAFTS 数据的应用，评估机器学习识别方法对 FAST 数据的整体性能表现。首先构建一个 CRAFTS 基准数据集；然后使用 PRESTO③ 对数据集预处理，包括去干扰、消色散和单脉冲搜索；再应用机器学习识别方法 (SPEGID④和 SPS⑤) 识别单脉冲搜索候选体；作为对比，我们也使用了启发式阈值判断方法 (RRATtrap⑥ 和 Clusterrank⑦)。为了全面比较不同机器学习分类

①SPEGID 特征 1−13 表征单个 SPEGs 信息，包括 SPEGs 峰值信噪比、脉冲宽度、DM 跨度、DM 与信噪比曲线对称值等信息，并提出一种新的峰值识别算法表征 DM 与信噪比曲线的峰值；特征 14−18 通过将一致 DM 范围内的 SPEGs 聚合为 SPEGs 组，统计 SPEGs 组内的最大信噪比、SPEGs 数量等信息，以反映一致 DM 范围内的 SPEGs 之间的关联[18]。

②SPS 特征 1−3 表征 SPEGs 的加权平均 DM 值、峰值信噪比以及对应的脉冲宽度；特征 4、5 通过超额峰度统计量，表征 SPEGs 的 DM 与信噪比和脉冲宽度曲线的峰值位置以及对称性的信息[19]。

③https://www.cv.nrao.edu/sransom/presto/

④https://github.com/dipangwvu/SPEGID

⑤https://github.com/danielemichilli/SpS

⑥https://github.com/ckarako/RRATtrap

⑦https://github.com/juliadeneva/clusterrank

器在识别方法中的性能差异，我们使用 Scikit-learn 库①实现的 7 种分类器：高斯朴素贝叶斯 (GaussianNB)[29]、逻辑回归 (LR)[30]、支持向量机 (SVM)[31]、决策树 (DT)[32]、随机森林 (RF)[33]、梯度提升决策树 (GBDT)[34] 以及多层感知机 (MLP)[35]；同时采用 SMOTE 技术为少数类 (脉冲星) 样本创建合成实例，以增加少数类的规模，构建平衡的数据集。

为了评估单脉冲搜索候选体识别方法的性能表现，需要构建一个基准数据集。我们构建的基准数据集包括 823 个脉冲星和 1 023 个非脉冲星候选体，其中脉冲星候选体是通过 RRATtrap 对 CRAFTS 数据初步筛选确认含有脉冲星信号的样本；非脉冲星候选体是在人工诊断中发现广泛造成数据产生虚假候选体的主要射频干扰类型构成的样本。此外，机器学习识别方法训练机器学习分类器须构建一个完全标注的特征数据集。对于基准数据集中候选体产生的每条特征数据，我们采用人工标注。标准如下：对于脉冲星候选体每条特征数据比对候选诊断图 (见图 2 a), b))，如果 SPEGs 对应的 DM 满足，1) 与时间空间的形状呈纺锤状 (见图 2 c))，2) 与信噪比曲线呈高斯曲线(见图 2 d))，3) 上下波动为 2 $\text{pc}\cdot\text{cm}^{-3}$ 范围内存在位于不同时间的 SPEGs，则被标注为脉冲星，否则标注为非脉冲星；对于非脉冲星候选体的每条特征数据，全部标注为非脉冲星。

3.1 SPEGID 方法对基准数据集的候选体自动识别与分类

为了实现 SPEGID 方法对 CRAFTS 基准数据集候选体的自动识别与分类，我们按照 SPEGID 方法收集每个候选体的特征数据②：先通过 DBSCAN 算法对候选体聚类得到 SPEGs；再根据 SPEGID 的特征工程收集每个 SPEGs 的特征值，共收集 227 632 条特征数据，每条数据对应一个 SPEGs；随后采用上述人工标注数据的标准标注所有数据，构建完全标注 SPEGID 特征数据集。最终，27 521 个 SPEGs 被标记为脉冲星，200 111 个 SPEGs 被标记为非脉冲星。为了研究不同机器学习分类器对 SPEGs 的分类性能，我们将 SPEGID 特征数据集分为训练集和测试集，通过训练集训练 7 种机器学习分类器。训练集包括 5 772 个脉冲星 SPEGs 和 87 898 个非脉冲星 SPEGs (对应 240 个脉冲星和 260 个非脉冲星候选体)；测试集为其余 21 749 个脉冲星 SPEGs 和 112 213 个非脉冲星 SPEGs (对应 583 个脉冲星和 763 个非脉冲星候选体)。针对训练集中脉冲星与非脉冲星样本不平衡问题，采用 SMOTE 合成脉冲星 SPEGs 样本，构建平衡的训练集，再应用 7 种机器学习分类器在平衡训练集进行训练。训练过程中，采用交叉验证方法，将训练集随机分为 5 组， 4 组用于训练分类器，1 组用于评估分类器；并通过 Scikit-learn 的网格搜索方法 (GridSearchCV) 确定每种分类器最佳的超参数值。在网格搜索中，我们为每种分类器定义一组超参数候选值，通过对所有可能的超参数组合进行评估，确定表现最佳的超参数组合作为分类器的最终超参数值③。最后，使用训练好的机器学习分类器对测试集进行识别与分类，并根据分类器结果和

①https://scikit-learn.org/stable/index.html

②SPEGID 方法收集特征数据时，DBSCAN 算法的超参数取值为：ϵ 邻域的距离阈值取 10，核心对象所需要的 ϵ 邻域的样本数取 12，其他经验参数与 SPEGID 方法保持一致。

③GaussianNB 使用默认的超参数。LR：正则化强度的倒数 C 取 18，采用拟牛顿法优化损失函数；SVM：目标函数的惩罚系数 C 取 10，核函数取 rbf，核函数系数 gamma 取 0.1；DT：树的最大深度取 12，叶子节点最小样本数取 8；RF：样本集切分策略采用 gini 指数，树的最大深度取 8，最多特征数取 4，叶子节点最小样本数取 4，树的颗数取 200；GBDT：树的

SPEGs 样本的人工标签计算性能评价指标。

表 2 列出了 7 种机器学习分类器对 SPEGID 测试集分类的评价指标值。可以看到：1) 所有采用 SMOTE 技术得到的平衡训练集训练的分类器具有更高的召回率和假阳性率，表明不平衡处理可以使更多脉冲星 SPEGs 正确分类，也会导致更多非脉冲星 SPEGs 被误判为脉冲星。2) 除了 SVM 出现过拟合 (SVM 在选择多组超参数值的情况下，召回率均低于 2%) 外，其他分类器各项性能指标值均在 80% 以上，这表明 SPEGID 开发的特征工程较好地实现了脉冲星与射频干扰的分离。3) 14 种分类器方法中，$\mathrm{LR_{smote}}$ 召回率最高，GBDT 和 LR 假阳性率最低 (SVM 分类器出现过拟合，其假阳性率不具有对比价值)；$\mathrm{GBDT_{smote}}$ 的 F1-score 和 G-mean 最高，表明 $\mathrm{GBDT_{smote}}$ 均衡性最好。总之，SPEGID 方法应用于基准数据集，即使选择简单分类器模型 (GaussianNB) 都可以取得较高评价指标值，这表明 SPEGID 能够较好地完成 CRAFTS 数据单脉冲搜索候选体识别任务。

表 2　7 种机器学习分类器对 SPEGID 测试集分类的性能数据

分类器	准确率	查准率	召回率	假阳性率	F1-score	G-mean
GaussianNB	0.944	0.760	0.955	0.058	0.847	0.948
LR	0.972	0.932	0.890	0.013	0.911	0.937
SVM*	0.839	0.968	0.012	0.001	0.025	0.114
DT	0.975	0.963	0.879	0.006	0.919	0.935
RF	0.990	0.988	0.950	0.002	0.968	0.974
GBDT	0.990	0.992	0.949	0.002	0.970	0.973
MLP	0.986	0.966	0.949	0.006	0.958	0.971
$\mathrm{GaussianNB_{smote}}$	0.933	0.719	0.963	0.072	0.824	0.945
$\mathrm{LR_{smote}}$	0.939	0.735	0.977	0.068	0.839	0.954
$\mathrm{SVM_{smote}}^*$	0.841	0.947	0.019	0.000	0.037	0.138
$\mathrm{DT_{smote}}$	0.980	0.977	0.900	0.004	0.937	0.947
$\mathrm{RF_{smote}}$	0.989	0.980	0.951	0.004	0.965	0.973
$\mathrm{GBDT_{smote}}$	0.990	0.984	0.960	0.003	0.972	0.978
$\mathrm{MLP_{smote}}$	0.987	0.959	0.959	0.008	0.959	0.975

注：分类器下标 smote 表示 SMOTE 技术平衡后的训练集训练的分类器；* 号表示 SVM 分类器出现过拟合，所有性能数据不具有对比价值。

3.2　SPS 方法对基准数据集的候选体自动识别与分类

为了使用 SPS 方法对 CRAFTS 基准数据集的候选体自动识别与分类，我们按照 SPS 方法收集每个候选体的特征数据①：先通过 Friends-of-Friends 算法对候选体聚类得到 SPEGs；再根据 SPS 的特征工程收集每个 SPEGs 的特征值，共收集 90 494 条数据；随后采用上述人工标注数据的标准对每条数据进行手工标注，构建完全标注 SPS 特征数据集。最终，14 779 个 SPEGs 被标记为脉冲星，75 715 个 SPEGs 被标记为非脉冲星。同样，我们将 SPS 特征数据集分为训练集和测试集。训练集包括 2 821 个脉冲星 SPEGs 和 37 896 个非脉冲星 SPEGs

最大深度取 5，最多特征数取 6，叶子节点最小样本数取 4，树的颗数取 50，子采样取 0.8；MLP 由输入层、隐藏层、输出层组成，激活函数为 relu，优化器选择 Adam。

① SPS 方法收集特征数据时，Friends-of-Friends 的超参数以及 SPS 方法使用的经验参数保持不变。由于 CRAFTS 数据由 FAST 单波束接收机收集，故未根据空间信息对标记为脉冲星的候选体做进一步处理。

(对应 240 个脉冲星和 260 个非脉冲星候选体)；测试集为其余 11 958 个脉冲星 SPEGs 和 37 819 个非脉冲星 SPEGs (对应 583 个脉冲星和 763 个非脉冲星候选体)。然后，我们用训练集训练 7 种机器学习分类器，并通过网格搜索方法 (GridSearchCV) 确定每种分类器最佳超参数值[①]；测试集评估分类器的性能表现。采用 SMOTE 数据不平衡处理技术合成脉冲星 SPEGs 样本，再应用 7 种机器学习分类器在平衡的训练集进行训练。最后使用训练好的机器学习分类器对测试集进行识别与分类；并根据每个分类器的分类结果与 SPEGs 样本人工标签，计算分类器的性能评价指标值。

表 3 总结了 7 种机器学习分类器对 SPS 测试集分类的评价指标值。可以看到：1) 相比于不平衡训练集训练的分类器，应用 SMOTE 技术得到的平衡训练集训练的分类器提高了召回率，也提高了假阳性率 (GaussianNB，LR 在选择多组超参数值的情况下，F1-score 值均低于 50%，对于二分类任务，这样的性能表现甚至不如随机猜测，表明 SPS 方法构造的特征训练这两种分类器效果不佳，它们假阳性率不具有对比价值)。2) 对于 MLP 分类器，$\mathrm{MLP}_{\mathrm{smote}}$ 召回率大幅提高了近 0.4。3) 在各项指标上，基于树的 3 种分类器 (DT, RF, GBDT) 优于另外 4 种分类器。4) 14 种不同分类器方法中，$\mathrm{GBDT}_{\mathrm{smote}}$ 具有最高的召回率 (如上所述，GaussianNB 分类器效果不佳，其召回率不具有对比价值)，GBDT 分类器具有最低的假阳性率；GBDT 的 F1-score 指标最高，$\mathrm{GBDT}_{\mathrm{smote}}$ 的 G-mean 指标最高。总之，SPS 方法应用于基准数据集，仅基于树的 3 种分类器 (DT, RF, GBDT) 取得较高评价指标值。

表 3　7 种机器学习分类器对 SPS 测试集分类的性能数据

分类器	准确率	查准率	召回率	假阳性率	F1-score	G-mean
GaussianNB*	0.253	0.243	0.999	0.983	0.391	0.129
LR*	0.557	0.329	0.813	0.523	0.468	0.622
SVM	0.846	0.875	0.417	0.019	0.565	0.639
DT	0.922	0.910	0.747	0.023	0.821	0.854
RF	0.933	0.944	0.765	0.014	0.845	0.868
GBDT	0.939	0.951	0.788	0.013	0.862	0.882
MLP	0.865	0.896	0.495	0.018	0.638	0.697
$\mathrm{GaussianNB}_{\mathrm{smote}}{}^*$	0.253	0.243	0.999	0.983	0.391	0.128
$\mathrm{LR}_{\mathrm{smote}}{}^*$	0.566	0.335	0.826	0.516	0.477	0.632
$\mathrm{SVM}_{\mathrm{smote}}$	0.858	0.703	0.711	0.095	0.707	0.802
$\mathrm{DT}_{\mathrm{smote}}$	0.889	0.727	0.863	0.102	0.789	0.880
$\mathrm{RF}_{\mathrm{smote}}$	0.902	0.759	0.863	0.086	0.808	0.888
$\mathrm{GBDT}_{\mathrm{smote}}$	0.885	0.699	0.920	0.125	0.795	0.897
$\mathrm{MLP}_{\mathrm{smote}}$	0.828	0.596	0.886	0.189	0.713	0.847

注：分类器下标 smote 表示 SMOTE 技术平衡后的训练集训练的分类器；* 号表示 GaussianNB 和 LR 分类器效果不佳，所有性能数据不具有对比价值。

[①] GaussianNB 使用默认的超参数；LR：正则化强度的倒数 C 取 5，采用拟牛顿法优化损失函数；SVM：目标函数的惩罚系数 C 取 10，核函数取 rbf，核函数系数 gamma 取 1；DT：叶子节点最小样本数取 10；RF：内部节点再分最小样本数取 2，叶子节点最小样本数取 2，树的颗数取 400；GBDT：叶子节点最小样本数取 10，树的颗数取 400，子采样取 0.9；MLP 由输入层、隐藏层、输出层组成，激活函数为 relu，优化器选择 Adam。

3.3 不同识别方法的性能表现和筛选速度的对比分析

为了对比不同类别识别方法在基准数据集中的性能表现，我们也采用了启发式阈值判断方法 (RRATtrap 和 Clusterrank) 对基准数据集进行候选体识别。由于启发式阈值判断的方法通过整体返回启发式分数表示候选体是脉冲星的概率，所以其分类结果依赖候选体[10, 11]。而机器学习识别方法分类结果依赖 SPEGs，且一个候选体通常包含多个 SPEGs (见图 2 a))。为了对比分析不同类别识别方法在候选体中的性能表现，对于 SPEGID 和 SPS 分类结果，我们规定每个候选体中，包含有评定为脉冲星的 SPEGs 被标记为脉冲星，否则标记为非脉冲星；再根据标注结果，计算 SPEGID 和 SPS 对于候选体的性能数据。值得注意的是，SPEGID 和 SPS 的实验结果是在完整特征数据集 (未划分训练集和测试集) 测试结果，以便公平地比较四种方法在基准数据集中的性能表现。对于启发式阈值判断 RRATtrap, Clusterrank，我们测试了多种阈值组合①。因为 F1-score 综合衡量了识别方法正确识别脉冲星以及产生虚假候选体的表现，我们选择启发式阈值判断方法具有最高 F1-score 值的阈值组合②对应的性能数据进行对比分析；对于机器学习识别方法，同样选择具有最高 F1-score 值的机器学习分类器③进行对比分析。

表 4 列出了 4 类识别方法对于候选体的性能数据，作为对比，最后两行汇总了 SPEGID 和 SPS 对于 SPEGs 的性能数据。可以看到：1) 4 类方法都表现出较高的召回率 (90% 以上)，意味着基准数据集中的绝大多数脉冲星信号被正确识别；同时，RRATtrap, Clusterrank, SPS 具有较高的假阳性率，预示着识别结果中包含大量的虚假候选体。Clusterrank 表现出最高的召回率 (97.7%) 和最高的假阳性率 (60.3%)。2) 用候选体衡量性能表现时，SPEGID 与 SPS 表现出与基于 SPEGs 结果相似的召回率。然而，基于候选体的假阳性率显著增加 (SPS 从 9.9% 到 46.1%，SPEGID 从 0.1% 到 4.7%)；这种假阳性率增加表明，被错误归类为脉冲星 SPEGs 分布在大量的候选体中。3) SPEGID 在多个指标取得最好的分数，且假阳性率仅有 4.7%，远远低于其他 3 种方法。同时，SPEGID 的 F1-score、G-mean 指标超过 95%，表明 SPEGID 取得了最好的性能数据 (高召回率、低假阳性率)。对于 CRAFTS 基准数据集，4 类识别方法具有相似的召回率，仅 SPEGID 方法取得相对较低的假阳性率，这主要是因为[17] SPEGID 方法开发的特征工程，能很好地挖掘出脉冲星在 DM 与时间和信噪比 (S/N) 空间独特的特征区分脉冲星与射频干扰。总之，对于 CRAFTS 基准数据集，SPEGID 取得了最好性能表现，显著优于其他 3 类方法。此外，通过对比表 1 和表 4 的结果发现，与其他射电望远镜相比 (见表 1)，CRAFTS 数据覆盖更广泛的频率范围 (270 ~ 1 620 MHz)，使用 4 类识别方法对 CRAFTS 数据识别单脉冲搜索候选体会表现出更高的假阳性率 (见表 4)，这表明 CRAFTS 数据可能会包含有大量类似于脉冲星信号的射频干扰数

①RRATtrap 取至少有 1 个打分为 (6, 5, 4, 3) 的组合以及打分为 6 的数量在 (2, 3, 4, 6) 以上，Clusterrank 对 R^2 值取 0.9 和 0.8、离群点判断阈值取 (0.5, 0.4, 0.25, 0.125)、是否开展 Bonferroni 校正进行组合。

②RRATtrap 阈值组合取打分为 6 的数量大于等于 3，Clusterrank 选择 R^2 值大于等于 0.9、离群点判断阈值取 0.25、不进行 Bonferroni 校正的阈值组合。

③SPEGID 选择进行 SMOTE 数据不平衡处理的 GBDT 分类器，SPS 选择未进行 SMOTE 数据不平衡处理的 GBDT 分类器。

据。针对 CRAFTS 数据的假阳性率较高的问题，建议未来收集 CRAFTS 数据中代表性的射频干扰样本，并进行针对性的时域/频域分析，挖掘消除射频干扰的有效特征，以减少虚假候选体对搜索结果的影响，从而降低人工诊断工作量和候选数据存储压力。

表 4　4 类识别方法对于候选体的性能数据

方法	准确率	查准率	召回率	假阳性率	F1-score	G-mean
RRATtrap	0.848	0.774	0.934	0.221	0.847	0.853
Clusterrank	0.657	0.568	0.977	0.603	0.718	0.622
SPEGID	0.953	0.949	0.954	0.047	0.951	0.953
SPS	0.717	0.612	0.948	0.461	0.744	0.715
$\mathrm{SPEGID}_{\mathrm{SPEGs}}$	0.994	0.991	0.959	0.001	0.975	0.979
$\mathrm{SPS}_{\mathrm{SPEGs}}$	0.903	0.643	0.914	0.099	0.755	0.907

注：下标 SPEGs 表示 2 类机器学习识别方法对于 SPEGs 的性能数据。

最后，我们研究了 4 类识别方法筛选单脉冲搜索候选体的运行速度①。一般来说，处理每个候选体的时间随候选体中单脉冲事件数 (SPE) 增加而增大。为了清楚显示 SPE 对识别方法筛选速度的影响，我们分别记录 4 类方法对基准数据集中脉冲星候选体筛选所花费的时间。值得注意的是，测试中 SPEGID 和 SPS 方法仅记录提取特征所花费的时间，并未统计人工标注特征数据集以及训练机器学习分类器所花费的时间。不同识别方法平均每小时筛选候选体数为 SPS: 4 010, SPEGID: 51, Clusterrank: 147, RRATtrap: 112，如图 3 所示，RRATtrap 与 SPEGID 时间受候选体中单脉冲事件数的影响很明显，而 Clusetrrank 和 SPS 时间并没有随事件数增加而发生明显的变化。总的来说，SPS 具有最快筛选速度，且时间差距随单脉冲事件数增加而显著增大，这个结果主要是因为 SPS 没有像另外三种方法一样构造峰值识别算法，检验 DM 与信噪比曲线是否存在峰值，导致其具有较快的速度。

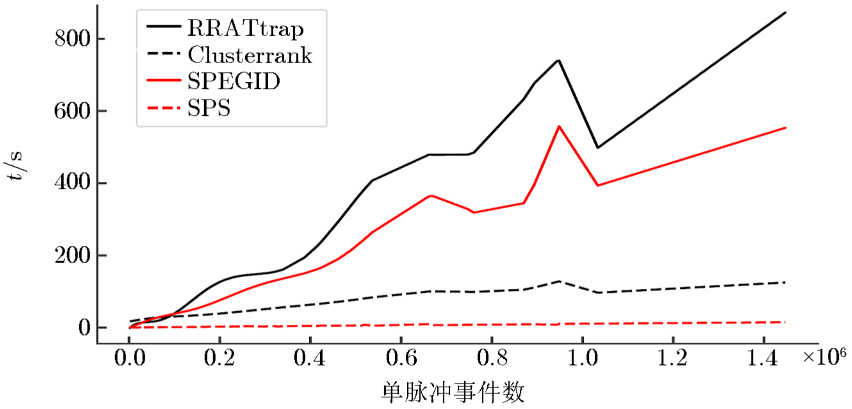


图 3　4 类识别方法随候选体中单脉冲事件数增加用时的趋势

① 本次测试硬件条件：Intel i7-6700 CPU、NVIDIA GTX 1080 Ti GPU、16 GB 内存，操作系统是 Ubuntu 18.04。

4 总结与讨论

本文研究了基于机器学习的单脉冲搜索候选体识别方法 (SPEGID, SPS) 应用到 CRAFTS 超宽带脉冲星巡天数据的性能表现以及筛选单脉冲搜索候选体的速度；作为对比，也使用了启发式阈值判断方法 (RRATtrap, Clusterrank)。首先构建一个基准数据集，包括 CRAFTS 数据初步筛选得到的脉冲星以及非脉冲星候选体；然后根据机器学习识别方法的特征工程，收集每个候选体特征值，逐条进行人工标注；并将特征数据分为训练集和测试集，用训练集训练 7 种机器学习分类器，测试集研究分类器的性能表现。同时针对训练集中脉冲星与非脉冲星样本不平衡问题，采用了 SMOTE 数据不平衡处理技术。结果表明 SPEGID 取得了最好性能表现 (高召回率、低假阳性率)，SPS 具有最快筛选速度。

根据对比分析结果，我们讨论未来如何对 FAST 观测数据开展高效的单脉冲搜索候选体识别。总之，单脉冲搜索候选体识别两种模式都是通过构建最大限度区分脉冲星与射频干扰的目标函数实现候选体的识别与分类。启发式阈值判断方法一般仅根据脉冲星在 DM 与时间和信噪比 (S/N) 空间的特点，构建启发式规则识别候选体，导致其仅能识别有限的脉冲星，且不能有效过滤射频干扰。Aggarwal 等人研究[36]发现：不正确的 DM 值、不正确的匹配滤波值、观测数据出现在主波束的位置等因素，往往会导致探测到的脉冲星信噪比出现损失①。同时，脉冲星信号在强度、宽度以及轮廓等方面会表现出显著差异[17]。因此，很难找到一套有效且普遍适用的启发式规则区分脉冲星与射频干扰。而机器学习识别方法通常在完全标注的特征数据集上进行训练，训练过程中同时学习了脉冲星和射频干扰的特点；所以，机器学习识别方法在识别脉冲星以及去除射频干扰方面均表现出良好的性能。此外，考虑到机器学习识别方法是一个循环迭代过程，包括分析分类结果、增加或修改数据、更新分类器、应用分类器以及重复上述步骤[37]，相信随着训练数据的积累，机器学习识别方法的性能会不断提高。因此，对于 FAST 观测数据中的单脉冲搜索候选体，建议选择机器学习识别方法自动识别与分类。

另外值得注意的是，最近深度学习也开始应用到单脉冲搜索候选体识别工作中。例如，Connor 和 Van Leeuwen[38]提出使用树状深度神经网络对单脉冲搜索候选体诊断图进行识别与分类；Agarwal 等人[39] 开发 FETCH 工具对 ASKAP 和 Parkes 数据的单脉冲搜索候选体诊断图进行实时分类；刘艳玲等人[40]通过卷积神经网络自动识别候选体诊断图，实现脉冲星与 FRB 分类。在后续的研究中，我们会进一步分析机器学习 (包括深度学习) 应用于 FAST 观测数据区别于其他射电望远镜观测数据的独有特征以及 CRAFTS 数据在候选体识别数据处理中的注意事项。

致谢

感谢审稿人对文章提出的评论意见和建议，使得文章质量有了显著的提高。本工作在 FAST (500 米口径球面射电望远镜) 数据基础上完成。FAST 是由中国科学院国家天文台运

① 单脉冲搜索候选体识别实践表明，信噪比低的脉冲星信号更容易被各种候选体识别方法错误识别[18]。

行和管理的国家大科学装置。

参考文献：

[1] Lorimer D R, Kramer M. Handbook of pulsar astronomy. Cambridge: Cambridge University Press, 2004: 6

[2] Larsson S. ApJS, 1996, 117: 197

[3] Cordes J M, Mclaughlin M A. ApJ, 2003, 596: 1142

[4] Mclaughlin M A, Lyne A G, Lorimer D R, et al. Nature, 2006, 439: 817

[5] Lorimer D R, Bailes M, McLaughlin M A, et al. Science, 2007, 318: 777

[6] Lyon R J, Stappers B W, Cooper S, et al. MNRAS, 2016, 459: 1104

[7] Deneva J S, Cordes J M, McLaughlin M A, et al. APJ, 2009, 703: 2259

[8] Keane E F, Ludovici D A, Eatough R P, et al. MNRAS, 2010, 401: 1057

[9] Burke-Spolaor S, Bailes M, Johnston S, et al. MNRAS, 2011, 416: 2465

[10] Karako-Argaman C, Kaspi V M, Lynch R S, et al. ApJ, 2015, 809: 67

[11] Deneva J S, Stovall K, McLaughlin M A, et al. ApJ, 2016, 821: 10

[12] Zhu W W, Berndsen A, Madsen E C, et al. ApJ, 2014, 781: 117

[13] 许余云, 李菂, 刘志杰, 等. 天文学进展, 2017, 35: 304

[14] Wang H F, Zhu W W, Guo P, et al. Science China: Physics, Mechanics & Astronomy, 2019, 62: 5

[15] Xiao J, Li X R, Lin H T, et al. MNRAS, 2020, 492: 2119

[16] Devine T R, Goseva-Popstojanova K, McLaughlin M. MNRAS, 2016, 459: 1519

[17] Pang D, Goseva-Popstojanova K, Devine T, et al. MNRAS, 2018, 480: 3302

[18] Pang D, Goseva-Popstojanova K, McLaughlin M, et al. PASP, 2020, 132: 4502

[19] Michilli D, Hessels J W T, Lyon R J, et al. MNRAS, 2018, 480: 3457

[20] Li D, Wang P, Qian L, et al. IEEE Microwave Magazine, 2018, 19: 112

[21] Lyon R J. Dissertation. Manchester: the University of Manchester, 2016: 15

[22] Yu Q Y, Pan Z C, Qian L, et al. RAA, 2020, 20: 91

[23] Ester M K, Hans P, Sander J, et al. kdd, 1996, 96: 226

[24] Huchra J P, Geller M J. ApJ, 1982, 257: 423

[25] Lyon L J, Brooke J M, Knowles J D, et al. International Conference on Pattern Recognition, 2014, 20: 1969

[26] 周志华. 机器学习. 北京: 清华大学出版社, 2016: 15

[27] Chawla N V, Japkowicz N, Kotcz A. ACM SIGKDD explorations newsletter, 2004, 6: 12

[28] Chawla N V, Bowyer K W, Lawrence O H, et al. Journal of artificial intelligence research, 2002, 16: 321

[29] Friedman N, Geiger D, Goldszmidt M. Machine learning, 1997, 29: 131

[30] Hosmer J, David W, Lemeshow S, et al. Applied logistic regression. New York: John Wiley & Sons, 2013: 36

[31] Hearst M A, Dumais S T, Osuna E, et al. IEEE Intelligent Systems and their applications, 1998, 13: 18

[32] Loh W Y. Wiley interdisciplinary reviews: data mining and knowledge discovery, 2011, 1: 14

[33] Breiman L. Machine learning, 2001, 45: 5

[34] Friedman J H. Annals of statistics, 2001, 29: 1189

[35] Svozil D, Kvasnicka V, Pospichal J. Chemometrics and intelligent laboratory systems, 1997, 39: 43

[36] Aggarwal K, Burke-Spolaor S, Law C J, et al. ApJ, 2021, 914: 53

[37] Domingos P. Communications of the ACM, 2012, 55: 78

[38] Connor L, Van Leeuwen J. AJ, 2018, 156: 256

[39] Agarwal D, Aggarwal K, Burke-Spolaor S, et al. MNRAS, 2020, 497: 1661

[40] 刘艳玲, 陈卯蒸, 李健, 等. 天文学报, 2022, 63: 107

Application of Single-Pulse Search Candidate Identification Based on Machine Learning to FAST Observation CRAFTS Data

ZHANG Bin[1,2,3,4], YOU Shan-ping[1,3,4], XIE Xiao-yao[1,3,4], YU Xu-hong[1,3,4], LIANG Nan[1,3,4]

(1. Key Laboratory of Information and Computing Science Guizhou Province/School of Cyber Science and Technology, Guizhou Normal University, Guiyang 550001, China; 2. School of Mathematical Sciences, Guizhou Normal University, Guiyang 550001, China; 3. NAOC-GZNU FAST Early Science Data Center, Guiyang 550001, China; 4. Joint Center for FAST Sciences Guizhou Normal University Node, Guiyang 550001, China)

Abstract: As a powerful tool for pulsar detection, single-pulse search plays an important role in detecting rotating radio transient sources and fast radio bursts. In order to quickly screen out the most valuable single-pulse search candidates from massive radio survey data, candidate identification has developed from early heuristic threshold judgment to automatic identification based on machine learning. For FAST observations, the performance of machine learning-based single-pulse search candidate identification applied to the commensal radio astronomy FAST survey (CRAFTS) ultra-wideband pulsar data was studied. In the evaluation process, two automatic recognition methods, single pulse event group recognition (SPEGID) and single pulse search device (SPS), were used to automatically identify the single-pulse search candidates generated by the CRAFTS benchmark dataset through seven different machine learning classifiers. For comparison, heuristic threshold judgment methods (RRATtrap and Clusterrank) are also used. The results showed that SPEGID had the best performance (highest F1-score 95.1%, next highest recall 95.4%, lowest false positive rate 4.7%), and SPS had the fastest screening speed (an average of 4 010 candidates per hour). By comparing the results of the analysis, how to carry out efficient work based on FAST observation data is discussed single-pulse search candidate identification.

Key words: single-pulse search; candidate identification; machine learning; pulsar; FAST; CRAFTS